

Prediction of PIK3CA mutation with gene expression

Jun Kang *

* Corresponding author: jkang.alien@gmail.com

Introduction

- Brevity
- Logic and clarity
- Clean typing
- The problem

– PIK3CA mutation in selecting drug

Targeted therapy becomes a standard treatment in many cancer patients. Much targeted therapy requires a test for a specific cancer genomic alteration, to treat the patients. Many direct tests for the genomic alteration have been developed and proved for their clinical utility to find which patients will be responding to the targeted therapy.

The machine learning approach has been actively researched to detect genomic alterations. Machine learning can build a prediction model from a large number of predictors such as radiomic features [1], pathology image [2] or gene expression data [3]. Because the most direct genomic test is more specific and sensitive than predictive models, the machine learning approach might have a limited role in clinical practice. However, the machine learning prediction can be a second-best when the direct test fails.

Prediction RAS pathway activation with gene expression data was done in the previous study. [5] They trained pan-cancer The Cancer Genome Atlas (TCGA) data with a supervised elastic net penalized logistic regression classifier with stochastic gradient descent. The performance of their model was 84% with an area under the receiver operating characteristic (AUROC) curve and 63% with an area under the precision-recall (AUPR) curve. The authors suggested their approach can be applied to other genomic alterations.

PIK3CA encodes the p110 α catalytic subunit of phosphatidylinositol 3'-kinase (PI3K). PI3K is a protein kinase that phosphorylates phosphatidylinositol 4,5-bisphosphate (PIP₂) to make phosphatidylinositol 3,4,5-trisphosphate (PIP₃). Phosphatase and tensin homolog (PTEN) changes PIP₂ to PIP₃ in contrast, PI3K. PIP₃ is a second messenger to activate protein kinase B (AKT) which is a serine/threonine-specific protein kinase. AKT inhibits apoptosis and promotes cell proliferation. [6]

Breast cancer with PIK3CA mutation has been approved to use PIK3CA inhibitor in hormone receptor-positive HER2 negative subtype. [7] The PIK3CA mutation is the second most driver mutation after TP53. The PIK3CA mutation is most frequently founded in endometrial carcinoma (45%), and followed by breast invasive carcinoma (24%), cervical squamous cell carcinoma, and endocervical adenocarcinoma (20%) and colon adenocarcinoma (16%).

We apply a supervised elastic net penalized logistic regression model in prediction PIK3CA mutation. The purpose of this study is to know this prediction model approach can be applied not only to RAS pathway activation but also PIK3CA mutation across many cancer types.

Materials and Methods

Dataset

We used the TCGA pan-cancer dataset. TCGA is a cancer genomic consortium that archives data of exome sequencing, gene expression, DNA methylation, protein expression, and clinical data of more than 10000 cancer samples across 33 common cancer types. The gene expression TCGA pan-cancer dataset is batch-corrected with normalization. The TCGA dataset is publically available. PIK3CA mutation data was got using cgdscr rpackage.[8] Gene expression data was got from the GDAC firehose using the RTCGAToolbox R package. [9]

10845 cases were available for both PIK3CA mutation and mRNA expression data. 5128 out of 20502 genes were included in the modeling process after filtering with median absolute deviation as described at the modeling process method. 33 cancer type dummy variables were included in predictor variables.

The target variable was the PIK3CA mutation status. The status of PIK3CA was considered as positive when the case has following PIK3CA variants which are the target variables of the thescreen PIK3CA RGQ PCR Kit; C420R, E542K, E545A, E545D, E545G, E545K, Q546E, Q546R, H1047L, H1047R, H1047Y. The thescreen PIK3CA RGQ PCR Kit was approved as a companion diagnostics to treat with PIK3CA inhibitor by the U.S. Food and Drug Administration. We split the three-quarters of the dataset for the trainset and one quarter for the test set.

Modeling process

To narrow down potential predictors, Genes with a large the median absolute deviation (more than third-quartiles) were selected. Yeo-Johnson transformation was done to correct skewness. Centering and scaling were done. All preprocessing was done using recipe r package. [10] Penalized logistic regression was applied to prediction modeling. 10-fold cross-validation with targe variable stratification was done over the hyperparameter grid: $\lambda \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$, $\alpha \{0.0, 0.25, 0.5, 0.75\}$. Lambda is penalty scaling parameter and alpha is mixing parameter of penalty function $((1 - \alpha)/2\|\beta\|_2^2 + \alpha\|\beta\|_1)$. [11]

Accessing model performance

Model performance was evaluated with the area under the receiver operating characteristic (ROC) and area under the ROC (AUROC) and precision-recall (PR) and area under the curve (AUPR). The prevalence of PIK3CA mutation is low. The low prevalence of PIK3CA results in an imbalanced dataset which makes the prediction difficult. The AUROC is optimistic in terms of performance. The AUPR is more informative than AUROC on imbalanced datasets. [12] The modeling process was done with tidymodels rpackage. [13]

Results

prevalence rate of PIK3CA mutation

The prevalence rate of PIK3CA was 0.11 in all cases. The PIK3CA prevalence rate of each cancer type varied. The median prevalence rate of PIK3CA of each cancer type was 0.03 (range 0-0.33) (Figure 1).

Selecting model and performance estimation

In 10-fold cross-validation, the model with $\lambda = 0.01$ and $\alpha = 1.0$ (Ridge regression) showed best performance in terms of AUROC. The final model was trained with the selected hyperparameters with the entire trainset.

The trainset AUROC was 0.93 and the testset AUROC was 0.84. The AUPR of trainset was 0.66 and the testset AUPR was 0.39. (Figure 1A)

Performance of each cancer type

Because the prevalence of PIK3CA mutation varies across the cancer type, the performance of each cancer type was investigated. The AUROC and AUPR were positively correlated between the train set and test set in cancer type subanalysis. (Figure 1B) The AUPR was high in cancer type with high PIK3CA mutation rate such as colon, breast, Uterus cancer types. The AUROC did not correlate with the PIK3CA mutation rate of each cancer type. (Figure 1C)

Important predictors

Figure 2 shows the top 30 important predictors. The coefficient is the parameter of the predictor which represents the effect of the predictor on prediction. IGF1R mRNA expression was the strongest negative predictor and PTEN was the strongest positive predictor. Both genes and PIK3CA are key players in the tyrosine kinase pathway. The cancer type was an important predictor. Some cancer types including uterine carcinosarcoma (UCS), bladder urothelial carcinoma (BLCA), pancreatic adenocarcinoma (PAAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) are the strongest predictors.

Discussion

- The main message answers the question and main supporting evidence

Our model showed good performance to predict the PIK3CA mutation of various cancer types. This result shows that the supervised elastic net penalized logistic regression model can be applied not only to the RAS activation pathway but also PIK3CA mutation. Both the RAS activation pathway and PIK3CA mutation are important and common cancer genomic alterations. They have a significant effect on gene expression in cancer cells. It might be a challenging prediction of genomic alterations that are infrequent or have a weak effect on gene expression.

- Critical assessment opinions on any shortcomings in study design

Prediction modeling from the TCGA pan-cancer dataset has limitations regarding data preprocessing. Methods for cancer gene expression have been developed for research. To statistical analysis, the gene expression data is processed between-sample

normalization to remove the batch effect. If the model has been trained from the
between-sample normalization, a new sample can not be exactly processed like trainset.
A model based on gene expression TCGA pan-cancer dataset has a limitation on
preprocessing. It is necessary for developing a preprocessing method that is independent
with a dataset to apply the gene expression data to the prediction model.

- limitations in methods
 - Case imbalance
 - flaws in analysis
 - validity of assumption
- Comparison with other studies where inconsistencies are discussed
- Evaluate the results - not the authors

Our PIK3CA prediction model performed better than the RAS activation prediction
model of the previous study in terms of both AUROC (0.84 vs 0.75) and AUPR (0.39 vs
0.24) on test set prediction. Our test set is corresponding to the samples initially
filtered from training. The target variable of our study is more specific than the
previous study. The specific important mutations can affect stronger downstream gene
expression than the broad events pathway activation. The previous study might be a
more difficult prediction problem.

Our model includes cancer type predictor and they are stronger than gene expression
data. The varying prevalence of PIC3CA mutation across cancer type might reason for
the strong cancer type predictor. If the cancer type was wrong or can not be
determined, our model performance can be poor.

Some significant gene expression predictors were closely related to PTEN and the
PI3K pathway. PTEN and IGFR1R are the strongest gene expression predictor which
have negative and positive predictive power. IGF1R is a tyrosine kinase receptor that
activates PI3K. [14] Insulin receptor substrate-2 (IRS2) is the adaptor protein of IGF1R.
[15] PTEN is an important regulator of PIP₃ by dephosphorylating PIP₃ in contrast
PI3K.[6]

There is some study trying to predict genomic alterations from gene expression data.
A study performed PIK3CA mutation prediction by a gene-expression signature which
is a sum of the average of the logarithmic gene expression. The model showed good
performance AUROC 0.71 in an independent test set. [4,16] Another study predicted
copy number alterations with gene expression using a multinomial logistic regression
model with least absolute shrinkage and selection operator (LASSO). The prediction of
the 1p/19q code was very good with an AUROC of 0.997. The gene-level prediction
was good with an AUROC 0.75. [17] A Hidden Markov Model Approach for Prediction
of Genomic Alterations from Gene Expression Profiling.[18] A logistic regression model
was used for MYCN gene amplification in neuroblastoma. [19]

- Conclusions comments on possible biological or clinical implications and
suggestions for further research.

Our study suggests that the prediction of genomic alteration with gene expression
data is possible with good performance. However, improvement of performance is
required for applying to clinical test and the standardization of generation processing of
gene expression data is also needed.

Figure legends

- Figure 1. prevalence rate of PIK3CA across cancer types The abbreviations of cancer types are explained in the S1 Appendix.
- Figure 2. Summary of modeling results
- (A) Left: receiver operating characteristic (ROC) curve right: precision-recall (PR) curve of trainset and test set. The horizontal green line is the PIK3CA mutation rate (0.11) (B) Correlation between trainset and test set of the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) among cancer type. The abbreviations are explained in the S1 Appendix. (C) Correlation between the PIK3CA mutation rate of area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR).
- Figure 3. Coefficients of model
- (A) Top 30 high coefficients of mRNA. (B) Coefficients of cancer types. The abbreviations of cancer types are explained in the S1 Appendix.

Supporting information

- S1 Appendix.
- S2 Figure.
- S1 Table.
- S2 Table.
- S3 Table.

References

1. Derle L, Fronheiser M, Lu L, Du S, Hayes W, Leung DK, et al. Identification of NonSmall Cell Lung Cancer Sensitive to Systemic Cancer Therapies Using Radiomics. Clin Cancer Res. American Association for Cancer Research; 2020;26: 2151–2162. doi:10.1158/1078-0432.CCR-19-2942
2. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from nonSmall cell lung cancer histopathology images using deep learning. Nature Medicine. Nature Publishing Group; 2018;24: 1559–1567. doi:10.1038/s41591-018-0177-5
3. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Reports. Elsevier; 2018;23: 172–180.e3. doi:10.1016/j.celrep.2018.03.046
4. Loi S, Haibe-Kains B, Majjaj S, Lallemand F, Durbecq V, Larsimont D, et al. PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptorPositive breast cancer. PNAS. National Academy of Sciences; 2010;107: 10208–10213. doi:10.1073/pnas.0907011107
5. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Reports. Elsevier; 2018;23: 172–180.e3. doi:10.1016/j.celrep.2018.03.046
6. Cantley LC. The Phosphoinositide 3-Kinase Pathway. Science. American Association for the Advancement of Science; 2002;296: 1655–1657. doi:10.1126/science.296.5573.1655

7. André F, Ciruelos E, Rubovszky G, Campone M, Loibl S, Rugo HS, et al. Alpelisib for PIK3CA-Mutated, Hormone ReceptorPositive Advanced Breast Cancer. *New England Journal of Medicine*. Massachusetts Medical Society; 2019;380: 1929–1940. doi:10.1056/NEJMoa1813904
8. Jacobsen A, Luna A. Cgdsr: R-based API for accessing the MSKCC cancer genomics data server (CGDS). 2019.
9. Samur MK. RTCGAToolbox: A new tool for exporting TCGA Firehose data. *PLoS One*. 2014;9(9):e106397.
10. Kuhn M, Wickham H. Recipes: Preprocessing tools to create design matrices. 2020.
11. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33: 1–22.
12. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. Public Library of Science; 10: e0118432. doi:10.1371/journal.pone.0118432
13. Kuhn M, Wickham H. Tidymodels: Easily install and load the 'tidymodels' packages. 2020.
14. LeRoith D, Roberts CT. The insulin-like growth factor system and cancer. *Cancer Letters*. 2003;195: 127–137. doi:10.1016/S0304-3835(03)00159-9
15. He W, Craparo A, Zhu Y, O'Neill TJ, Wang L-M, Pierce JH, et al. Interaction of Insulin Receptor Substrate-2 (IRS-2) with the Insulin and Insulin-like Growth Factor I Receptors EVIDENCE FOR TWO DISTINCT PHOSPHOTYROSINE-DEPENDENT INTERACTION DOMAINS WITHIN IRS-2. *J Biol Chem*. American Society for Biochemistry and Molecular Biology; 1996;271: 11641–11645. doi:10.1074/jbc.271.20.11641
16. Cizkova M, Cizeron-Clairac G, Vacher S, Susini A, Andrieu C, Lidereau R, et al. Gene Expression Profiling Reveals New Aspects of PIK3CA Mutation in ERalpha-Positive Breast Cancer: Major Implication of the Wnt Signaling Pathway. *PLOS ONE*. Public Library of Science; 5: e15647. doi:10.1371/journal.pone.0015647
17. Mu Q, Wang J. CNAPE: A Machine Learning Method for Copy Number Alteration Prediction from Gene Expression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2019; 1–1. doi:10.1109/TCBB.2019.2944827
18. Geng H, Ali HH, Chan WC. A Hidden Markov Model Approach for Prediction of Genomic Alterations from Gene Expression Profiling. In: Măndoiu I, Sunderraman R, Zelikovsky A, editors. *Bioinformatics Research and Applications*. Berlin, Heidelberg: Springer; 2008. pp. 414–425. doi:10.1007/978-3-540-79450-9_38
19. He X, Qin C, Zhao Y, Zou L, Zhao H, Cheng C. Gene signatures associated with genomic aberrations predict prognosis in neuroblastoma. *Cancer Communications*. 2020;40: 105–118. doi:10.1002/cac2.12016